ANA CAROLINA
Cientista de Dados na TakeBlip

ESTATÍSTICA DESCRITIVA

AGENDA

- **Bloco 1: Estatística e tipos de variáveis**
- **Bloco 2: Medidas - resumo (Posição)**
  Intervalo longo - 10 min
- **Bloco 3: Medidas de Dispersão**
- **Bloco 4: Representações gráficas e distribuições**

NOSSO ENCONTRO DE HOJE

ALGUMA EXPECTATIVA ESPECÍFICA SOBRE A AULA?

Por que estamos aqui?

Perguntas, dúvidas e incertezas.

Dados (pesquisa, internet, experiência anteriores).

ESTATÍSTICA

É a ciência que corresponde a coletar, analisar, apresentar e interpretar dados, bem como tomar decisões com base nessas análises”.

- **Prem S. Mann**

Por que estamos aqui?

Perguntas, dúvidas e incertezas.

Dados (pesquisa, internet, experiência anteriores).

Estatística

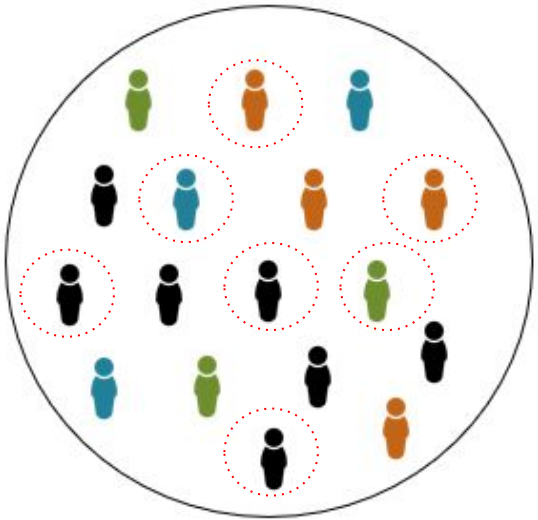
Informação

Direcionamento.

População x Amostra

Generalização (Inferência Estatística)

(Resultados e conclusões)



Amostragem

População

Conjunto de todos os elementos de interesse.

Amostra

Subconjunto da população, na qual conseguimos coletar e analisar os dados.

Estatística Inferencial

O objetivo é produzir afirmações sobre dada característica de uma população, na qual estamos interessados, a partir de informações colhidas de uma parte dessa população.

Estatística Descritiva

Aplicação de técnicas para descrever e resumir um conjunto de dados.

ESTATÍSTICA DESCRITIVA

Estatística Descritiva

É em geral, usada na etapa inicial de uma análise para:

- Descrever
- Sumarizar
- Fornecer uma visão geral

De modo que possamos tirar conclusões iniciais e obter um direcionamento a respeito das características de interesse.

Estatística Descritiva

As técnicas usadas na Estatística Descritiva são:

- **Gráficos descritivos:** Usamos vários tipos de gráficos para mostrar e sumarizar os dados. Por exemplo: Gráficos de barras, gráficos de pizza e etc.

- **Tabelas:** Usamos tabelas para exibir os dados. Por exemplo: Tabelas de frequência.

- **Medidas de resumo:** Resumimos os dados em números que expressam algumas de suas características. Por exemplo: Média, moda, mediana.

Exemplo:

Queremos comparar os salário dos Cientistas de Dados do Brasil em 2019 e 2020.

2019:
R$ 7500,00

2020:
R$ 8650,00

Vantagens

- Simplifica uma grande quantidade de dados
- Facilita comparações (entre grupos, pessoas)
- Auxilia tomadas de decisão

Desvantagem

- Implica em perda de informação

TIPOS DE VARIÁVEIS

O que são variáveis?

É a característica de interesse que é medida ou observada em cada indivíduo da amostra ou população.

Ex: Em um questionário, pergunta-se:

- Qual é a sua idade?
- Quantas moram com você?
- Qual é a renda total da sua família?
- Você tem emprego fixo?
- Qual é o seu estado civil?

Variáveis:

Idade	Moradores	Renda familiar	Emprego Fixo	Estado Civil
27	2	R$2000	Sim	Solteira
25	5	R$ 3500	Sim	Casado
32	1	R$ 1650	Não	Divorciado

Tipos de variáveis

Quantitativas

Consiste em um dado **numérico** que representa contagem ou medida. (Não há intervalos entre os valores possíveis.)

Exemplos:

- número de filhos (0, 1, 2, 3, ...)
- peso
- Idade
- Renda

Qualitativas

(Categórica ou de atributos) consiste em uma **característica** ou atributo que não possui um valor numérico. (Há intervalos entre os valores possíveis.)

Exemplos:

- sexo (masculino, feminino)
- cor dos olhos
- escolaridade (primário, médio, superior)
- grau de obesidade (leve, moderado, grave, mórbida)

Tipos de variáveis

Quantitativas

(Número)

Discreta

Geralmente se referem a contagens, e assumem apenas valores inteiros.

Exemplo:

- número de filhos (0, 1, 2, 3, ...)

Contínua

Geralmente se referem a medições e podem assumir valores fracionados.

Exemplo:

- peso (kg)
- altura (m)
- IMC (Kg/m2)

Qualitativas

(Característica)

Nominal

Não há uma ordem definida entre as categorias.

Exemplo:

- cor dos olhos (preto, azul, verde)

Ordinal

Há uma ordem definida entre as categorias.

Exemplo:

- escolaridade (primário, médio, superior)
- grau de obesidade (leve, moderado, grave, mórbida)

Mão na Massa

Classifique as seguintes variáveis em: qualitativa (nominal ou ordinal) ou quantitativa (discreta ou contínua).

(a) O tipo de câncer de pele: melanoma, carcinoma basocelular, carcinoma epidermóide e etc.
(b) O grau de dificuldade de um trilha de caminhada: leve, moderada, forte ou difícil.
(c) O número de larvas do mosquito da dengue em um recipiente.
(d) A quantidade de chuva em um dia (em milímetros).
(e) Frequência ao médico: semanal, mensal, trimestral.
(f) A cor da pele de pessoas (ex.: branca, negra, amarela).
(g) O número de consultas médicas feitas por ano por um associado de certo plano de saúde.
(h) A idade de uma pessoa, segundo as faixas etárias: 0 a 5 anos, 5 a 10 anos, ..., 100 anos ou mais.

Mão na Massa (Gabarito)

Classifique as seguintes variáveis em: qualitativa (nominal ou ordinal) ou quantitativa (discreta ou contínua).

(a) O tipo de câncer de pele: melanoma, carcinoma basocelular, carcinoma epidermóide e etc. Qualitativa nominal.
(b) O grau de dificuldade de um trilha de caminhada: leve, moderada, forte ou difícil. Qualitativa ordinal.
(c) O número de larvas do mosquito da dengue em um recipiente. Quantitativa discreta.
(d) A quantidade de chuva em um dia (em milímetros). Quantitativa contínua.
(e) Frequência ao médico: semanal, mensal, trimestral. Qualitativa ordinal.
(f) A cor da pele de pessoas (ex.: branca, negra, amarela). Qualitativa nominal.
(g) O número de consultas médicas feitas por ano por um associado de certo plano de saúde. Quantitativa discreta.
h) A idade de uma pessoa, segundo as faixas etárias: 0 a 5 anos, 5 a 10 anos, ..., 100 anos ou mais. Qualitativa ordinal.
i) O teor de gordura corporal de uma pessoa, medido em gramas por 24 horas. Quantitativa contínua.

MEDIDAS-RESUMO

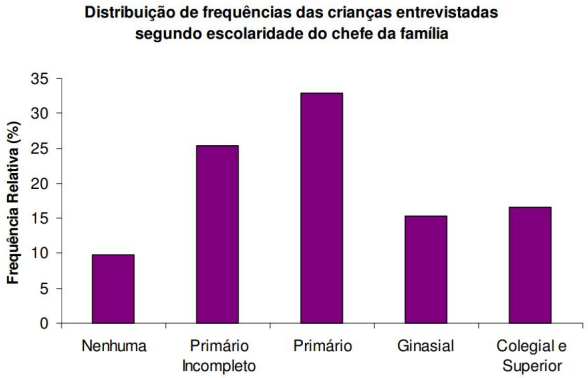
Como resumir os dados de uma variável qualitativa?

- **Frequências: absoluta e relativa**

Exemplo: Estudo das condições de saúde de crianças de um município brasileiro (Monteiro e Benício, 1987).

Distribuição de freqüências das crianças entrevistadas segundo escolaridade do chefe de família

Escolaridade do chefe da família	Freqüência Absoluta	Freqüência Relativa (%)	Freqüência Relativa Acumulada (%)
Nenhuma	99	9.8	9.8
Primário Incompleto	256	25.4	35.2
Primário	331	32.9	68.1
Ginasial	154	15.3	83.4
Colegial e Superior	167	16.6	100
Total	**1007**	**100**	------



Como resumir os dados de uma variável quantitativa?

Identificam o elemento (ou elementos) típico do conjunto de dados.

- **Média, Mediana, Moda**

Quantificam a dispersão (variabilidade) dos dados em torno do elemento típico.

- **Variância, desvio-padrão, amplitude, percentis**

Média Aritmética Simples (Média)

Cálculo: Soma de todos os valores dividida pela quantidade de total de dados.

Exemplo: Suponha que a turma de Data Science tenha estudantes com as idades descritas na tabela abaixo. Calcule a idade média dos estudantes dessa turma.

Idade (anos)
17
21
18
23
19
18
24

$$\bar{X} = \frac{17 + 21 + 18 + 23 + 19 + 18 + 24}{7} = \frac{140}{7} = 20$$

- A idade média dos estudantes da Turma de Data Science é de 20 anos.

Mediana

É o valor central em um conjunto de dados organizado de forma crescente.

É o valor que divide o conjunto de dados em dois grupos: os valores menores e os maiores que ele.

Para calcular a mediana, basta seguir o passo a passo:

1. Coloque os valores do conjunto de dados em ordem crescente.
2. Se a quantidade de valores do conjunto for **ímpar**, a mediana é o valor central.
3. Se a quantidade de valores do conjunto for **par**, basta somar os valores centrais e dividir por 2.

Mediana

Exemplo: Calcule a idade mediana dos estudantes da turma de Data Science.

Idade (anos)
17
21
18
23
19
18
24

→

Idade (anos)
17
18
18
19
21
23
24

Dados ordenados.

- Temos um conjunto de dados ímpar (7 valores), portanto a mediana é o valor central. A mediana da idade dos alunos de Data Science é de 19 anos.

Idade (anos)
17
18
18
19
21
23
24
27

→ $\frac{19+21}{2} = \frac{40}{2} = 20$

- Temos um conjunto de dados par (8 valores), portanto a mediana é a soma dos dois valores centrais divididos por 2.

Moda

É o valor ou categoria mais frequente em um conjunto de dados.

Dessa forma, para encontrá-la basta observar a frequência com que os valores aparecem.

Exemplo: Calcule a idade **mais frequente** dos alunos da turma de Data Science.

Idade (anos)
17
21
18
23
19
18
24

Idade (anos)
17
18
18
19
21
23
24

Dados ordenados.

O valor mais frequente ou a moda da idades dos alunos da turma de Data Science é de 18 anos.

Média, Moda e Mediana - Qual eu uso?

Observação sobre a média:

- A média é muito afetada pelos **valores presentes no conjunto de dados**, então, se tivermos valores muito **discrepantes** nesse conjunto, a média vai ser "puxada" para mais próximo desses valores.

Exemplo:

Idade (anos)
17
21
18
23
19
18
24

$$\bar{X} = \frac{140}{7} = 20$$

Vamos agora **acrescentar uma idade discrepante** a esse conjunto de dados.

Idade (anos)
17
21
18
23
19
18
24
81

$$\bar{X} = \frac{221}{8} = 27.6$$

Perceba que a média foi de 20 para 28 anos anos aproximadamente, **por causa de um único valor** (a idade discrepante de 81 anos).

Observação sobre a moda:

- Como a moda é determinada pela quantidade de vezes que um valor aparece nos dados, **os valores que pouco se repetem não vão afetá-la.**
- Isso pode ser um problema quando valores discrepantes nos dados passam despercebidos e impactam em uma análise mais aprofundada.

Exemplo:

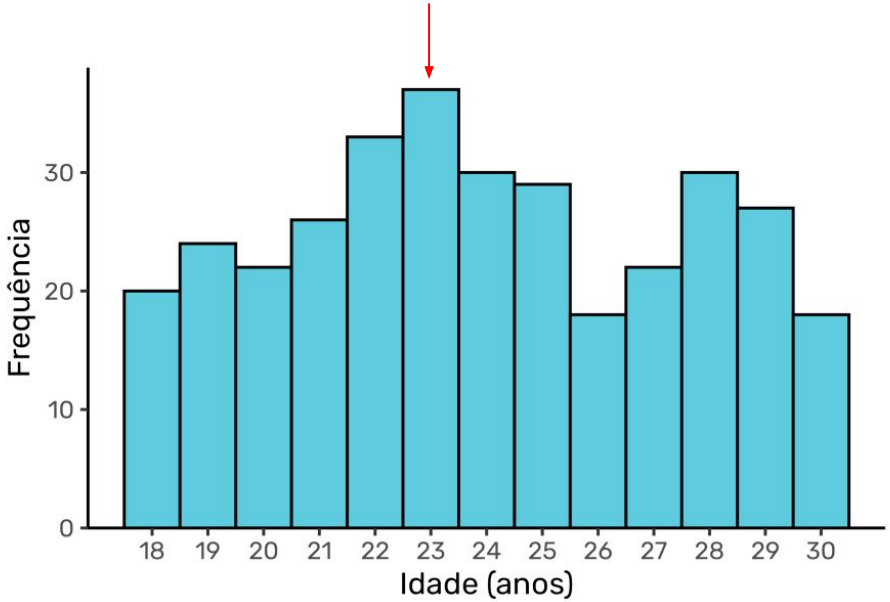
Idade (anos)
17
18
18
19
21
23
24

Idade (anos)
17
18
18
19
21
23
24
81

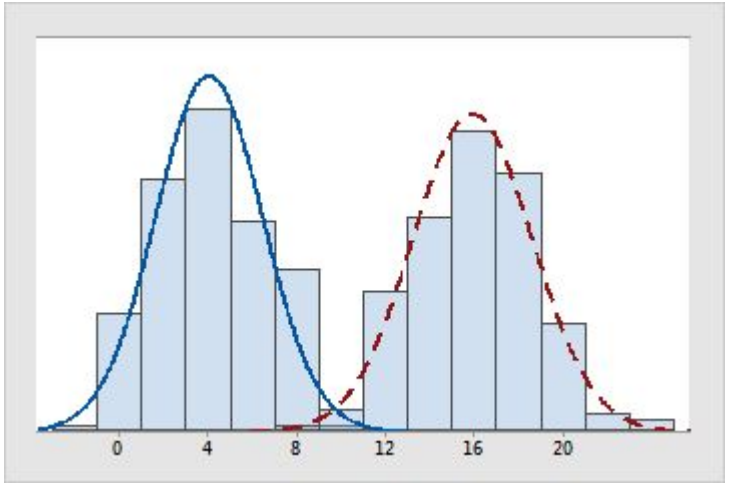
Para os dois casos (com e sem valor discrepante) a moda é igual a 18.

- Se no conjunto de dados não existir nenhum valor que se repete, dizemos que esse conjunto é amodal, ou seja, que não possui moda.
- Também podem existir conjuntos bimodais, isto é, com duas modas e assim por diante.

Qual a idade mais frequente entre os motoristas que acionam o seguro?



Distribuição Bimodal



Quando devemos usar a Moda então?

- Casos em que é útil saber qual é o valor ou valores mais frequentes nos dados.
- Quando queremos aplicar uma medida de resumo a dados nominais, isto é, quando não há valores numéricos.

Por exemplo, qual a nota mais frequente de um determinado grupo de estudantes.

Notas	Quantidade de alunos
A	3
B	10
C	7
D	5
E	4

A nota modal é B, porque é a nota com maior frequência na amostra.

Exemplo

Idade (anos)	Moda	Média
17, 21,18, 23, 19, 18, 24	18	20
17, 21,18, 23, 19, 18, 24 ,81	18	28

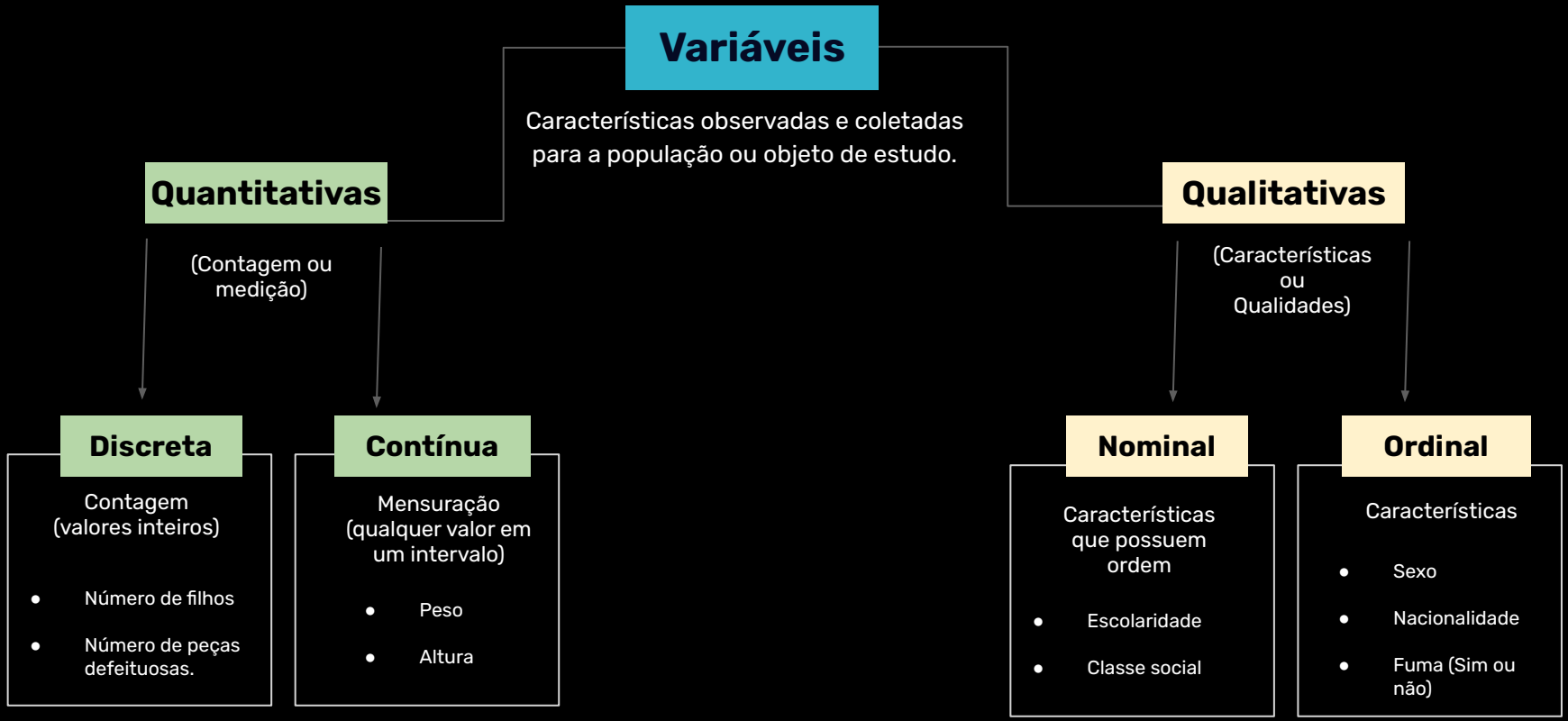
Então a gente tem uma medida que é muito influenciada pelos dados discrepantes (média) e outra que é pouco influenciada (moda), certo?

- E é aí que a **mediana** entra, como ela é calculada pelo valor central dos dados, ela é uma medida robusta em relação aos valores discrepantes.

Idade (anos)	Moda	Média	Mediana
17, 21,18, 23, 19, 18, 24	18	20	19
17, 21,18, 23, 19, 18, 24 ,81	18	28	20

ALGUMA DÚVIDA ATÉ AQUI?

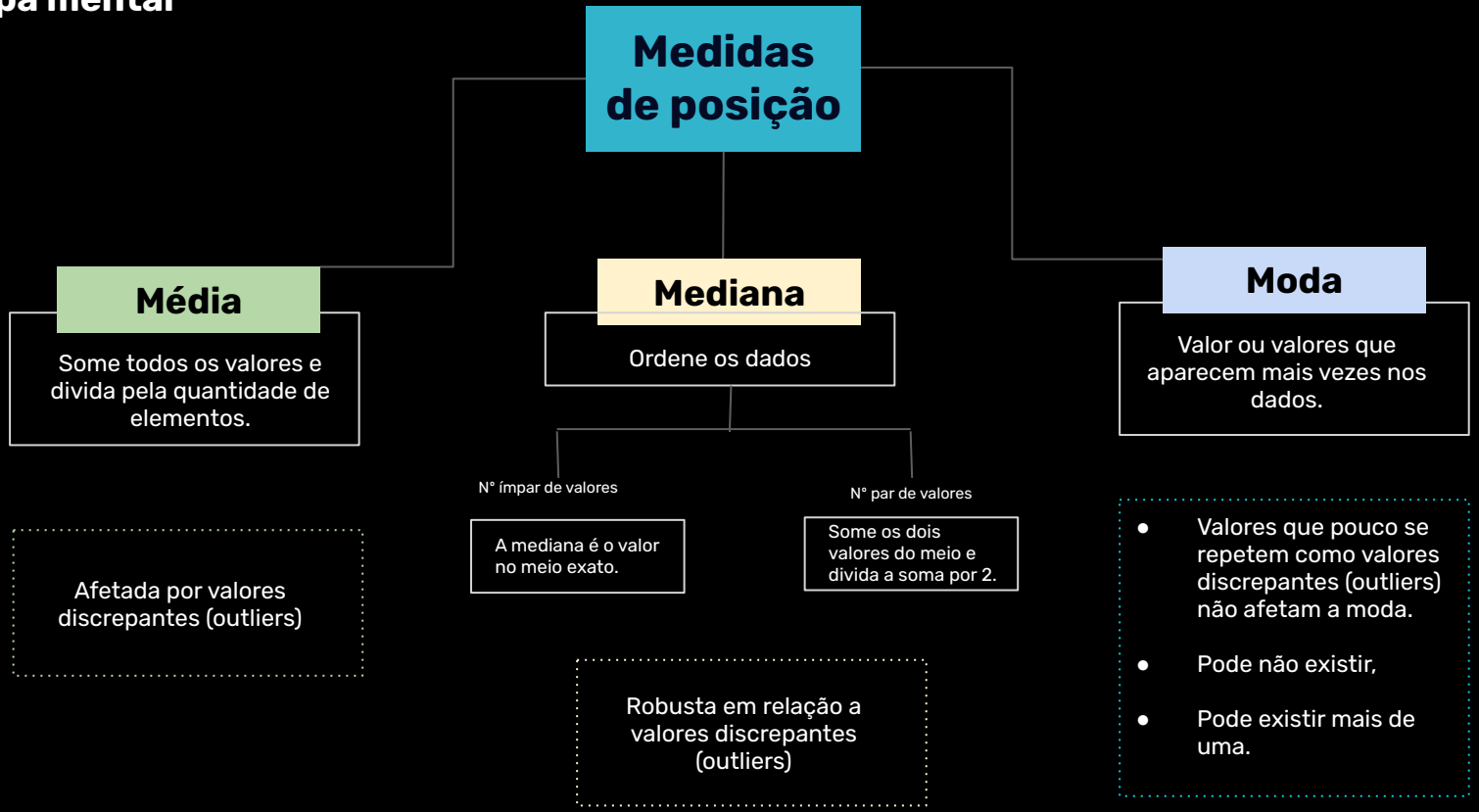
PARA LEMBRAR - Mapa mental



PARA LEMBRAR

Mapa mental

Identificam o elemento (ou elementos) típico do conjunto de dados.



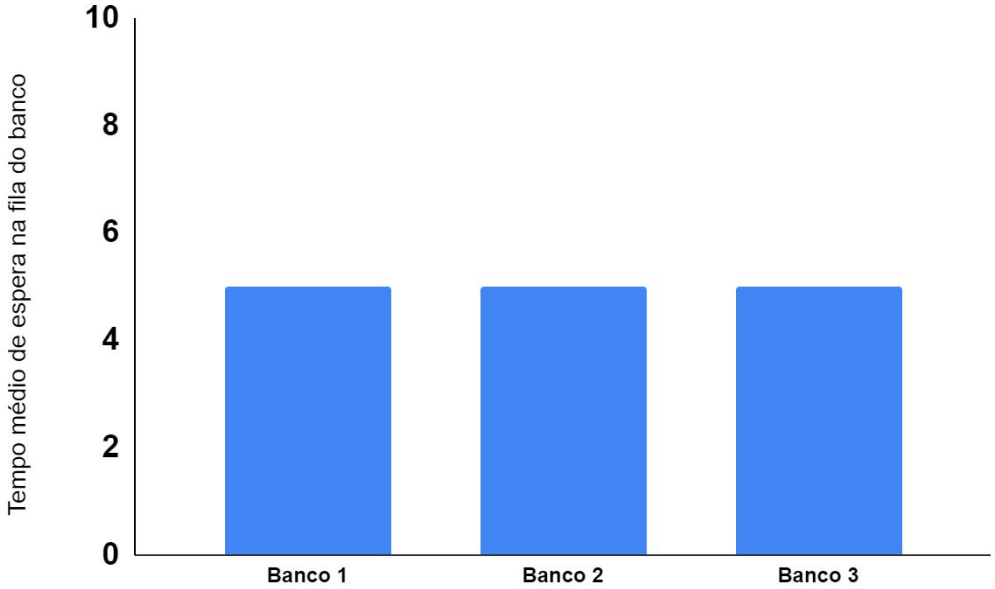
INTERVALO 10 MIN

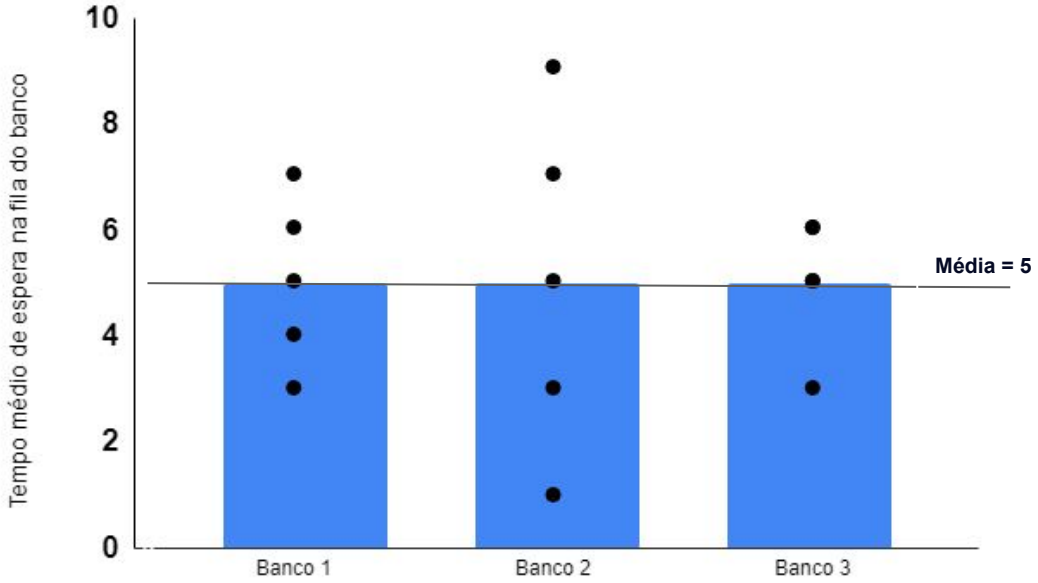
APROVEITE PARA:

- Fazer anotações do que viu até agora (aprendizados, insights, dúvidas)
- Levantar-se, esticar os braços e as pernas, relaxar por mais tempo
- Comer algo para voltar com energia renovada
- Ir ao toalete

Medidas de Dispersão

Qual a nossa conclusão a partir desse gráfico?





	Tempo de espera na fila em minutos				
	Cliente 1	Cliente 2	Cliente 3	Cliente 4	Cliente 5
Banco 1	3	4	5	6	7
Banco 2	1	3	5	7	9
Banco 3	3	5	5	6	6

Amplitude

A Amplitude de um conjunto de dados é a diferença entre o maior valor e o menor valor.

Amplitude = (Valor máximo - valor mínimo).

Exemplo: Considere os dados de tempo de espera de clientes em três bancos diferentes.

	Tempo de espera na fila em minutos				
	Cliente 1	Cliente 2	Cliente 3	Cliente 4	Cliente 5
Banco 1	3	4	5	6	7
Banco 2	1	3	5	7	9
Banco 3	3	5	5	6	6

Amplitude

- Banco 1: 7 - 3 = 4 min.
- Banco 2: 9 - 1 = 8 min.
- Banco 3: 6 - 3 = 3 min.

Exemplo: Em qual desses dois bancos o tempo de espera varia mais?

Tempo de espera na fila em minutos		
	Banco 1	Banco 2
Cliente 1	1	1
Cliente 2	3	5
Cliente 3	5	5
Cliente 4	7	5
Cliente 5	9	9

~~Amplitude~~

- Banco 1: 9 - 1 = 8 min.
- Banco 2: 9 - 1 = 8 min.

Variância

A variância é uma medida de dispersão que mostra o quão distante cada valor desse conjunto está do valor central (média).

Quanto menor é a variância, *mais próximos os valores estão da média*; mas quanto maior ela é, mais distantes os valores estão da média.

Exemplo:

Tempo de espera na fila em minutos	
	Banco 1
Cliente 1	1
Cliente 2	3
Cliente 3	5
Cliente 4	7
Cliente 5	9

Cálculo:

Tempo de espera na fila em minutos	
	Banco 1
Cliente 1	1
Cliente 2	3
Cliente 3	5
Cliente 4	7
Cliente 5	9

- Passo 1: Calcular a média. $\bar{X} = \frac{1+3+5+7+9}{5} = 5$

- Passo 2: Subtrair a média de cada um dos valores.

1 - 5	= -4
3 - 5	= -2
5 - 5	= 0
7 - 5	= 2
9 - 5	= 4

- Passo 3: Elevar os valores da diferença ao quadrado.

= -4	x^2 →	= 16
= -2		= 4
= 0		= 0
= 2		= 4
= 4		= 16

- Passo 4: Somar a diferença ao quadrado. $16 + 4 + 0 + 4 + 16 = 40$

Cálculo:

$$\text{Variância (Var)} = \frac{\text{Soma dos quadrados}}{n-1}$$

- n = quantidade de observações da amostra.

Tempo de espera na fila em minutos	
	Banco 1
Cliente 1	1
Cliente 2	3
Cliente 3	5
Cliente 4	7
Cliente 5	9

- Passo 5: Dividir a soma pelo número de observações - 1.

$$\text{Var} = \frac{40}{5-1} = \frac{40}{4} = 10$$

- Variância = 10 minutos²

Interpretação não intuitiva e difícil de interpretar.

Exemplo: Em qual desses dois bancos o tempo de espera varia mais?

Tempo de espera na fila em minutos		
	Banco 1	Banco 2
Cliente 1	1	1
Cliente 2	3	5
Cliente 3	5	5
Cliente 4	7	5
Cliente 5	9	9

Variância

- **Banco 1:** 10 minutos²
- **Banco 2:** 8 minutos²

Variância

$$s^2 = \frac{\sum_i^n (X_i - \bar{X})^2}{(n-1)}$$ → Variância **amostral.**

$$\sigma^2 = \frac{\sum_i^n (X_i - \bar{X})^2}{n}$$ → Variância **populacional.**

Desvio padrão

Corresponde à raiz quadrada da variância.
Está na mesma unidade de medida que os dados originais

$$s = \sqrt{s^2}$$

Desvio padrão **amostral**

$$\sigma = \sqrt{\sigma^2}$$

Desvio padrão **populacional**

Tempo de espera na fila em minutos		
	Banco 1	Banco 2
Cliente 1	1	1
Cliente 2	3	5
Cliente 3	5	5
Cliente 4	7	5
Cliente 5	9	9

Banco 1

s^2 = 10 minutos²
s = 3,16 minutos

Banco 2

s^2 = 8 minutos²
s = 2,8 minutos

Quanto maior o desvio-padrão, mais dispersos os dados estão em relação a média.

Observação sobre o desvio padrão:

- O desvio padrão é uma medida da variação de todos os valores a partir da média dos dados.
- O valor do desvio padrão é usualmente positivo. (Nunca é negativo). É zero quando todos os valores dos dados são o mesmo número.
- Quanto maior o valor do desvio padrão, maior a variação dos dados.
- O valor desvio padrão pode aumentar drasticamente com a inclusão de valores discrepantes (outliers) ou seja,valores de dados que estão afastados dos demais.
- A unidade do desvio padrão é a mesma unidade dos dados originais.

Coeficiente de variação

Como o desvio padrão é expresso na mesma unidade dos dados observados em estudo, **comparar dois ou mais conjuntos de valores que estão em unidades de medida diferentes torna-se impossível**.

O coeficiente de variação (CV) fornece a variação dos dados em relação à média.

$$\text{Cv} = \frac{s}{\bar{x}} * 100$$

$\text{Cv} = \text{Coeficiente de variação}$

$\text{s} = \text{Desvio padrão amostral}$

$\bar{x} = \text{Média amostral}$

Quanto menor for o valor do coeficiente de variação, mais homogêneos serão os dados, ou seja, menor será a dispersão em torno da média.

Coeficiente de variação

Idade (anos)
17
21
18
23
19
18
24

$$\bar{x} = 20 \text{ anos}$$

$$s = 2{,}7 \text{ anos}$$

$$\text{Cv} = \frac{s}{\bar{x}} * 100$$

$$\text{Cv} = \frac{2{,}7 \text{ anos}}{20 \text{ anos}} * 100 = 13,5$$

Altura (m)

$$\bar{x} = 1{,}67 \text{ m}$$

$$s = 0{,}5 \text{ m}$$

$$\text{Cv} = \frac{0{,}5 \text{ m}}{1{,}67 \text{ m}} * 100 = 29,94$$

ALGUMA DÚVIDA ATÉ AQUI?

Quartis

Idade (anos)
17
18
18
19
21
23
24

Sabemos que a mediana de um conjunto de dados é o valor do meio, de modo que 50% dos valores são iguais ou menores do que a mediana, e 50% dos valores são iguais ou maiores do que a mediana.

Assim como a mediana divide os dados em duas partes iguais, os três quartis, representados por Q_1 , Q_2 e Q_3, **dividem os valores ordenados em quatro partes iguais**.

Para obtenção dos quartis, o conjunto de dados deve ser organizado de forma crescente.

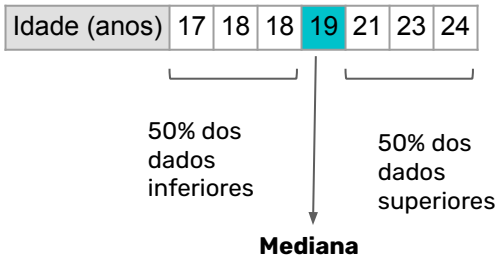
Exemplo: A idade dos estudantes da turma de Data Science.

Idade (anos)	17	21	18	23	19	18	24

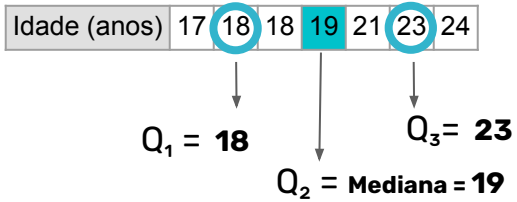
- Passo 1: Ordene os dados

Idade (anos)	17	18	18	19	21	23	24

- Passo 2: Encontre a mediana.

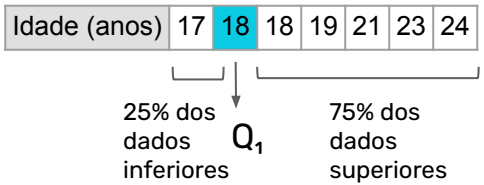


- Passo 3: Dividir as duas metades (inferiores e superiores no meio).



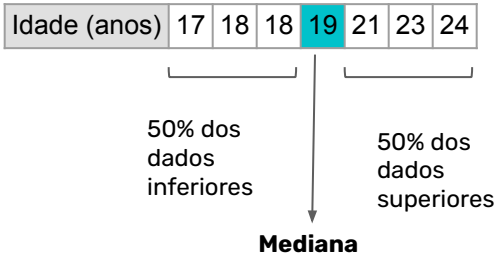
Q_1 (1º Quartil)

Separa os 25% dos valores ordenados inferiores dos 75% superiores.
(No mínimo 25% dos valores ordenados são menores ou iguais a Q_1, e no mínimo 75% dos valores são maiores ou iguais a Q_1 .



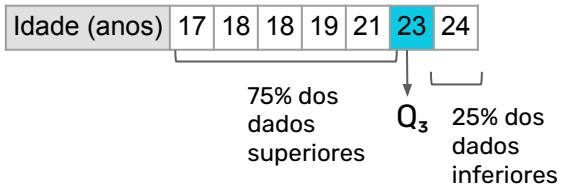
Q_2 (2° Quartil) - Mediana

O mesmo que a mediana, separa os 50% inferiores valores ordenados dos 50% superiores.



Q_3 (3° Quartil)

Separa os 75% dos valores ordenados inferiores dos 25% superiores. (No mínimo 75% dos valores ordenados são menores ou iguais a Q_3 , e no mínimo 25% dos valores são maiores ou iguais a Q_3.



Exemplo

Idade (anos)	15	16	18	18	19	21	22	24	24

Q_1 = **(16 + 18) / 2 = 17**

Q_3 = **(22+ 24) / 2 = 23**

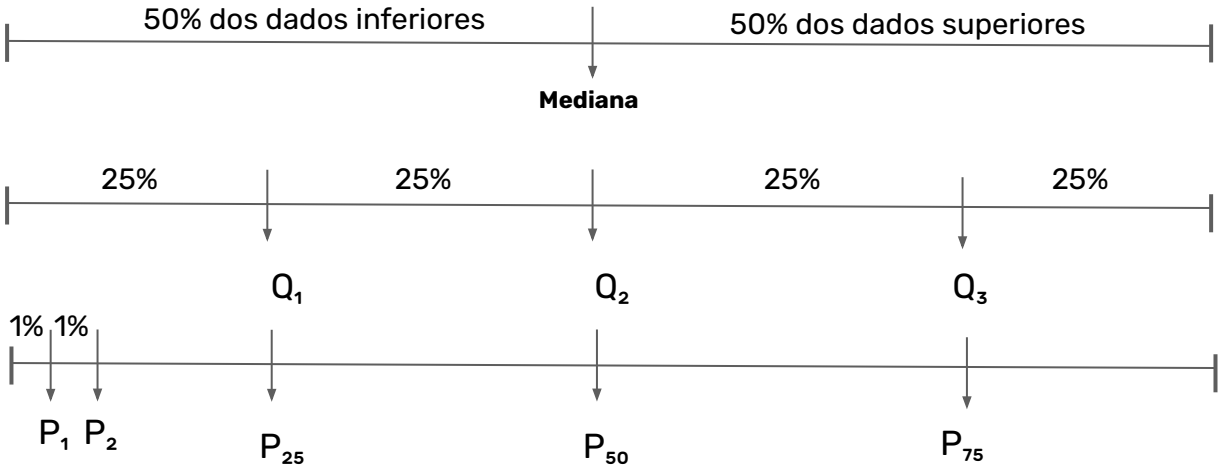
Para os casos em que não temos um número ímpar de valores para pegar o valor do meio, fazemos uma média dos valores.

Percentis

Assim como há 3 quartis que separam os dados em quatro partes, existe também os percentis, que dividem os dados em 100 partes.

Temos 99 percentis, representados por P_1 , P_2 , ..., P_{100}, que dividem os dados em 100 grupos com cerca de 1% dos valores em cada um.

$Q_1 = P_{25}$
$Q_2 = P_{50}$
$Q_3 = P_{75}$



Mão na Massa

Calcule as estatísticas descritivas das idades dos candidatos a uma vaga em uma fintech, listadas abaixo:

{23, 21, 27, 22, 20, 25, 19, 24, 24, 23, 22, 20, 26, 20, 18, 21, 20, 19, 21, 22, 28}

a) média
b) mediana
c) moda
d) desvio-padrão
e) amplitude
f) percentil 25
g) percentil 75
h) percentil 50

Mão na Massa (Gabarito)

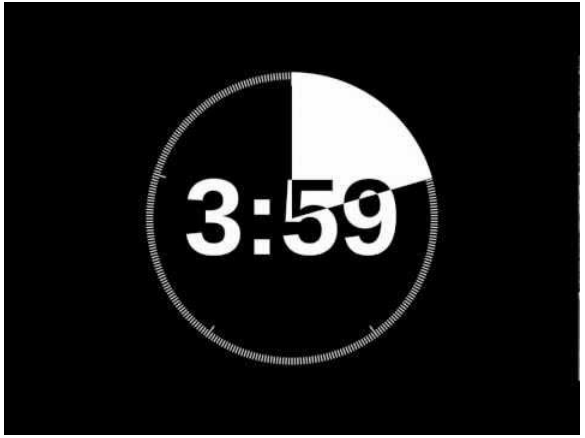
Calcule as estatísticas descritivas das idades dos candidatos a uma vaga em uma fintech, listadas abaixo:

{23, 21, 27, 22, 20, 25, 19, 24, 24, 23, 22, 20, 26, 20, 18, 21, 20, 19, 21, 22, 28}

a) média = 22,14 anos
b) mediana = 22 anos
c) moda = 20 anos
d) desvio-padrão = 2,73 anos
e) amplitude = 28 – 18 = 10 anos
f) percentil 25 = 20 anos
g) percentil 50 = 22 anos
h) percentil 75 = 24 anos

- Não se esqueça de ordenar a amostra para calcular algumas das estatísticas.

INTERVALO 5 MIN



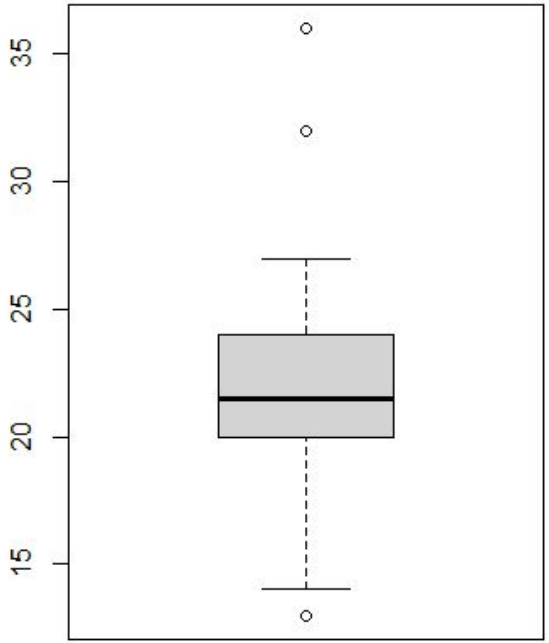
APROVEITE PARA:

- Fazer anotações do que viu até agora (aprendizados, insights, dúvidas)
- Levantar-se, esticar os braços e as pernas
- Ir ao toalete
- Pegar uma água, um chá, um snack

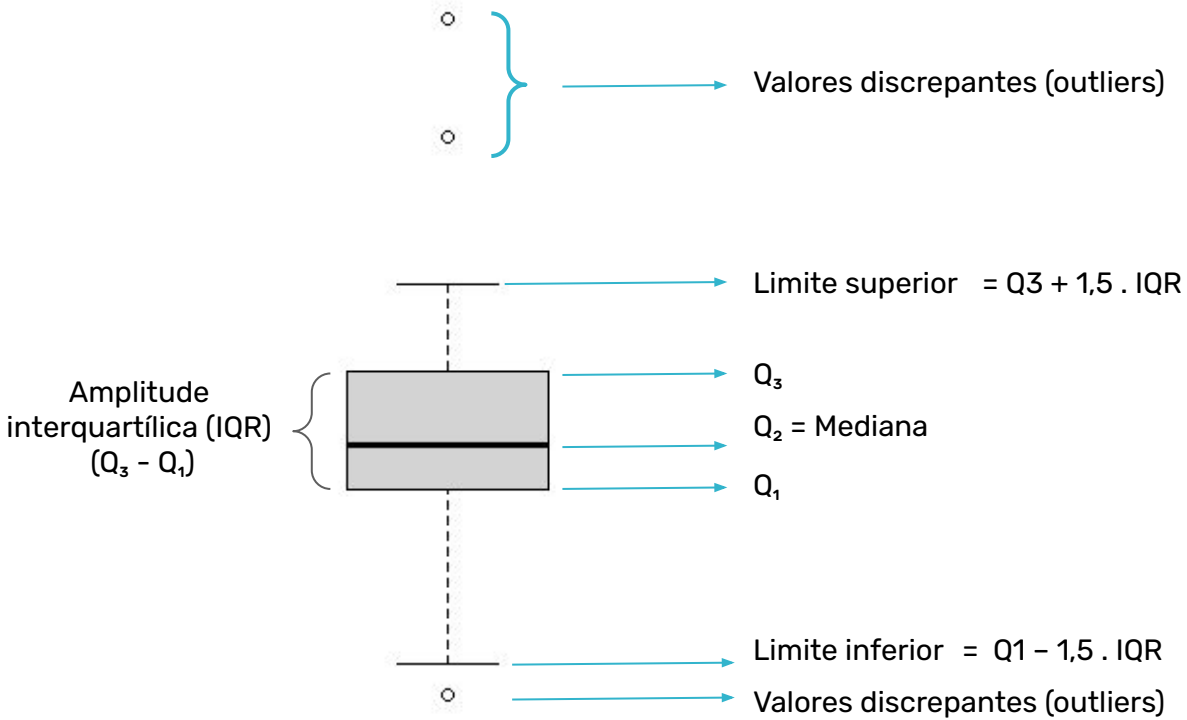
Representações gráficas

Boxplot

O boxplot é uma ferramenta gráfica para representar a variação (distribuição) de dados observados de uma variável numérica.



Boxplot


Valores discrepantes (outliers)
Limite superior = Q3 + 1,5 . IQR
Amplitude interquartílica (IQR) (Q_3 - Q_1)
Q_3
Q_2 = Mediana
Q_1
Limite inferior = Q1 – 1,5 . IQR
Valores discrepantes (outliers)

Boxplot

Exemplo:

{13, 14 ,18, 19, 19, 20, 20, 20, 20, 21, 21, 21, 22, 22, 23, 23, 24, 24, 25, 26, 27, 32, 36}

Q_1 = **20**

Q_2 = **Mediana = 21**

Q_3= **24**

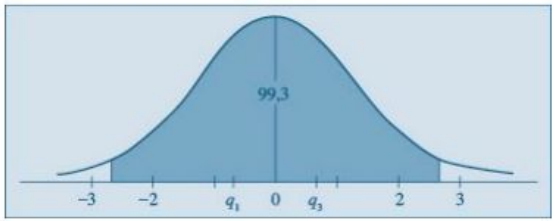
Amplitude interquartílica (IQR)
$(Q_3 - Q_1) = (24-20) = 4$

LIMITES TEÓRICOS

Limite inferior = Q1 – 1,5 . IQR
(20 - 1.5*4) = 14

Limite superior = Q3 + 1,5 . IQR
(24 + 1.5*4) = ~~30~~

Não existe o valor 30 na base de dados (o limite superior é o valor anterior ao 30 que tem na base = 27)



Exemplo:

{13, 14 ,18, 19, 19, 20, 20, 20, 20, 21, 21, 21, 22, 22, 23, 23, 24, 24, 25, 26, 27, 32, 36}

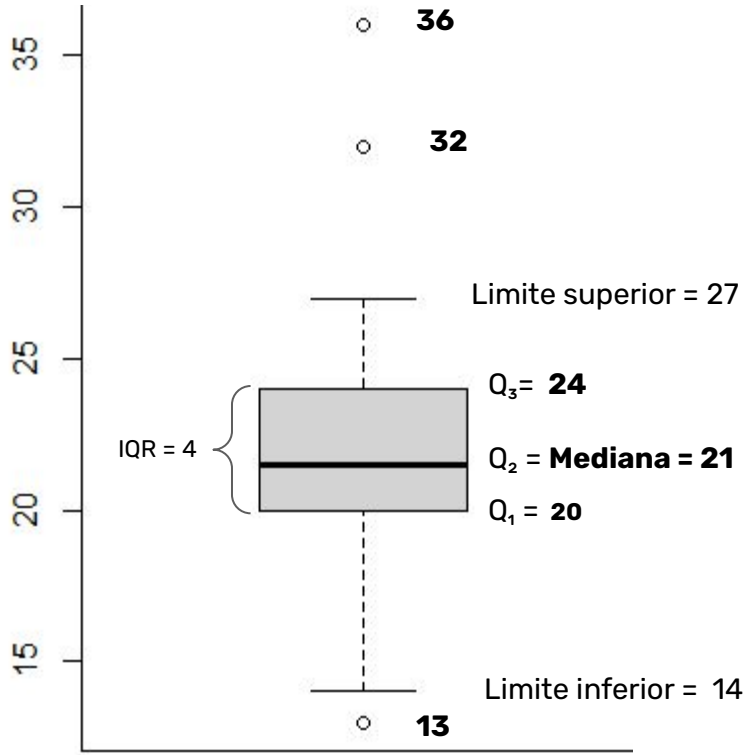


Gráfico de barras

Um gráfico de barras é uma forma de resumir um conjunto de dados categóricos ou discretos, em que contamos quantas vezes cada categoria ou valor aparece nos dados.

Figura 2.2 Gráfico em barras para a variável Y: grau de instrução.

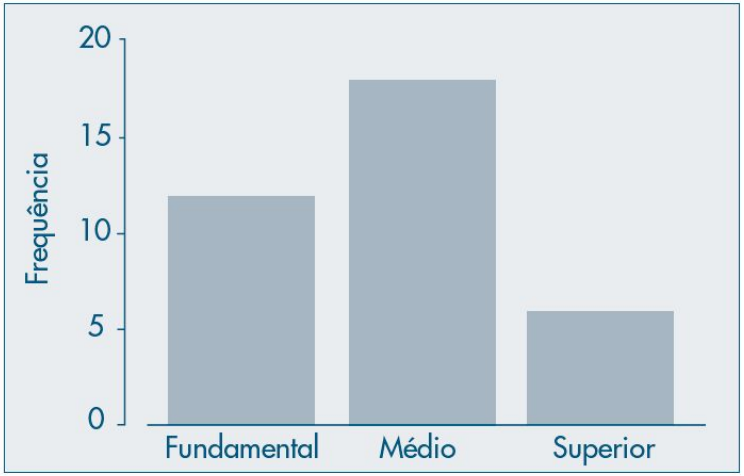
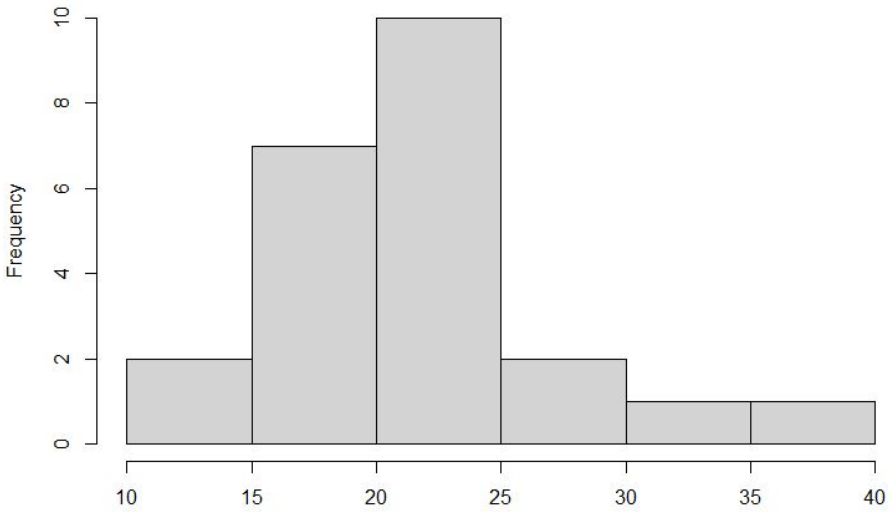


Figura 2.4 Gráfico em barras para a variável Z: número de filhos.

Histograma

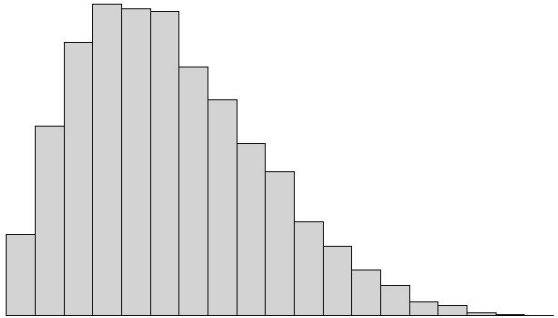
Exemplo:
{13, 14 ,18, 19, 19, 20, 20, 20, 20, 21, 21, 21, 22, 22, 23, 23, 24, 24, 25, 26, 27, 32, 36}



- **Centralidade**: qual é o centro da distribuição? Onde está a maioria das observações?
- **Amplitude**: os dados contém observações entre quais valores? Qual é o ponto de máximo e o ponto de mínimo?
- **Simetria**: temos a mesma frequência de observações com valor alto e com valor baixo? Será que a distribuição dos dados é simétrica ou valores mais altos são mais raros?

Assimetria (skew)

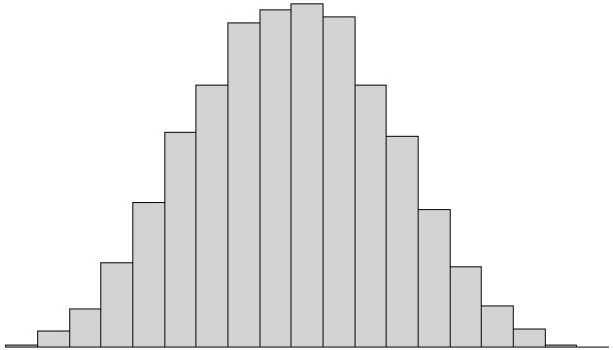
As distribuições apresentadas nos histogramas podem se organizar de maneiras diferentes, e podemos classificá-los em diferentes tipos:



Distribuição com assimetria positiva

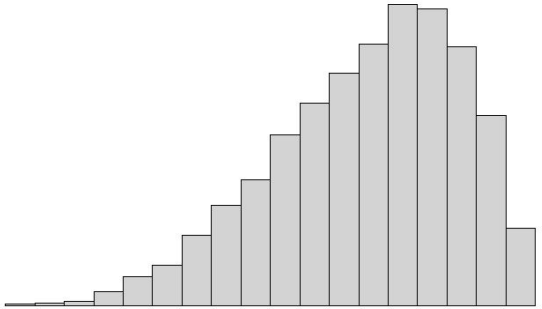
Apresenta maior concentração de dados nos valores mais baixos. **A cauda mais longa da distribuição fica à direita, indicando a ocorrência de valores altos com baixa frequência.**

Esse tipo de distribuição é denominada assimétrica positiva ou à direita, sendo bastante comum em administração e economia: variáveis como preços, PIB, salários, etc., possuem, em geral, este comportamento.



Distribuição simétrica

Apresenta frequência mais alta no centro e ir diminuindo de acordo com aproximação das bordas, tanto a da direita, quanto a da esquerda, **apresentando um formato de sino**.



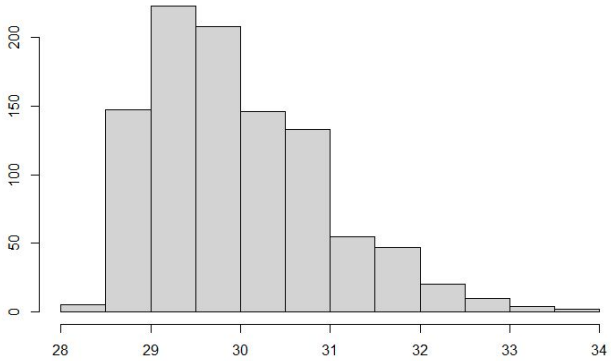
Distribuição com assimetria negativa

Apresenta maior concentração de dados nos valores mais altos. **A cauda mais longa da distribuição fica à esquerda, indicando a ocorrência de valores pequenos com baixa frequência**.

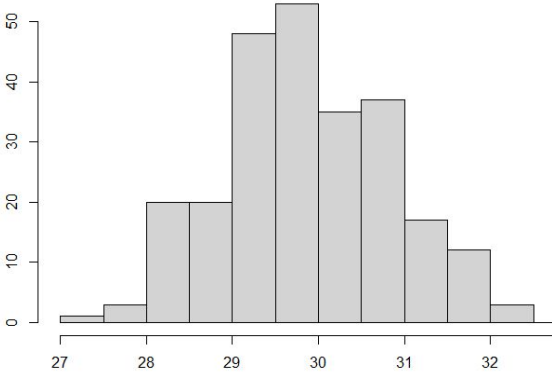
Esse tipo de distribuição é denominada assimétrica negativa ou à esquerda.

Assimetria (skew)

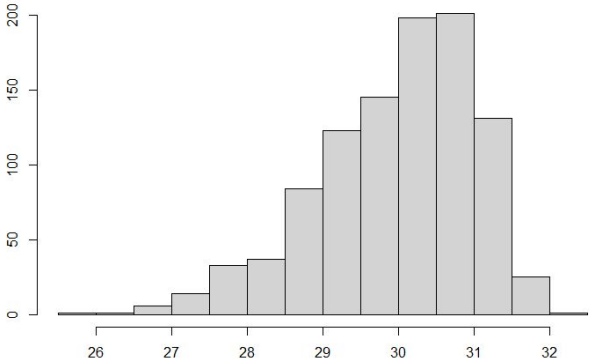
Idade (anos)



Distribuição com assimetria positiva



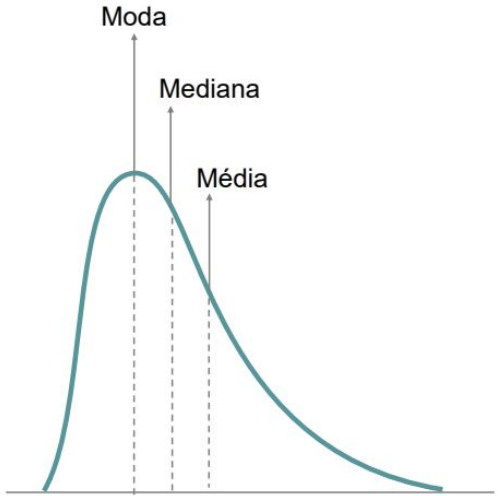
Distribuição simétrica



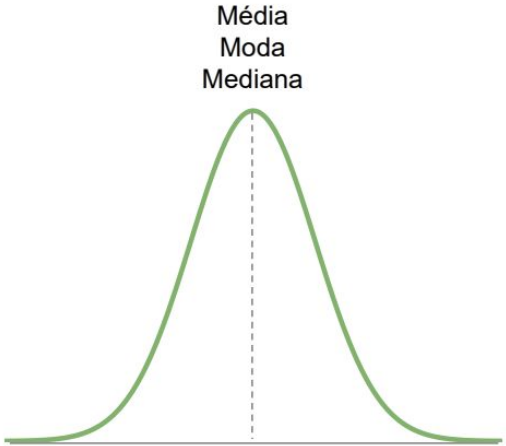
Distribuição com assimetria negativa

Para distribuições simétricas, a média é uma boa medida descritiva para o conjunto de dados.

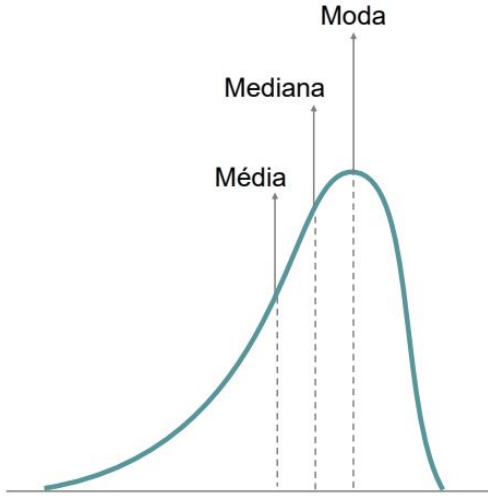
Assimetria (skew)



Distribuição com assimetria positiva



Distribuição simétrica



Distribuição com assimetria negativa

Exemplo

Vamos a um exemplo prático: Suponha que estamos avaliando o retorno de dois investimentos que fizemos em ações.

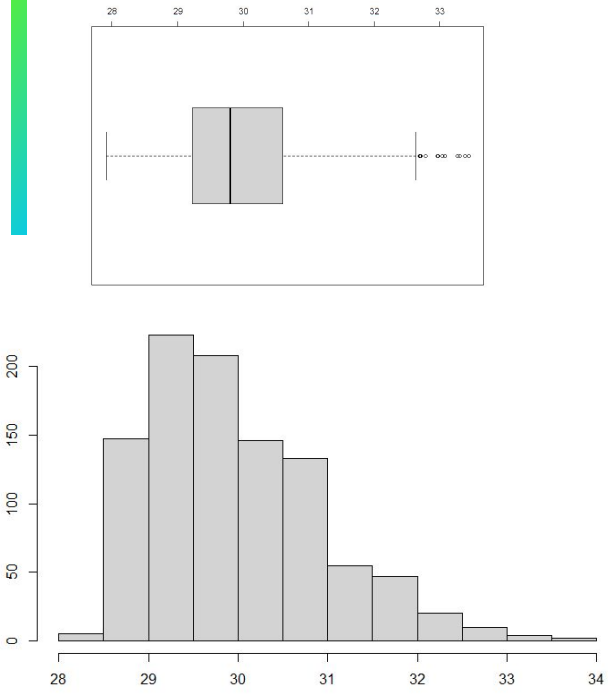
Sabemos que ambos têm a mesma média e mesma variância, no entanto, um deles possui assimetria positiva e o outro assimetria negativa.

- Um investimento com assimetria negativa apresenta uma cauda longa e fina de valores à esquerda da média (induzindo a possibilidade de perdas elevadas). Ou seja nesse caso, na maior parte do tempo você tem um retorno alto, mas existe a possibilidade de ter perdas muito grandes.

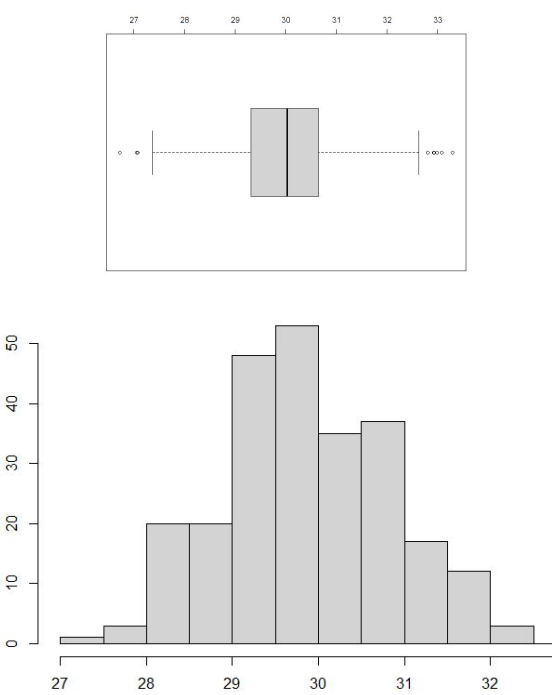
- Por outro lado, assimetria positiva é sinônima de cauda longa e fina de valores à direita da média (induzindo a possibilidade de elevados ganhos). Ou seja nesse caso, na maior parte do tempo você tem um retorno baixo, mas existe a possibilidade de ter um retorno muito alto.

Essa informação pode ser útil, caso o investidor tenha que se decidir por uma dessas aplicações.

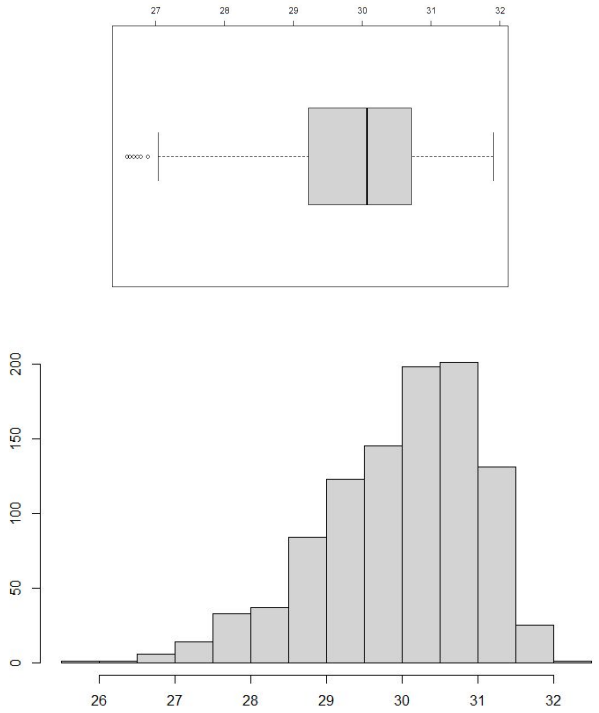
Assimetria (skew)



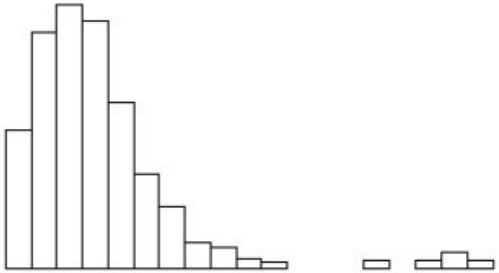
Distribuição com assimetria positiva

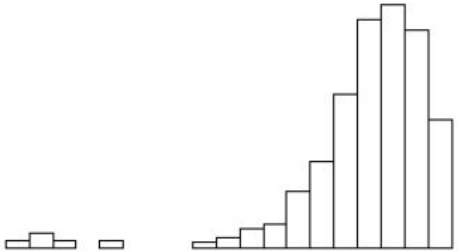


Distribuição simétrica



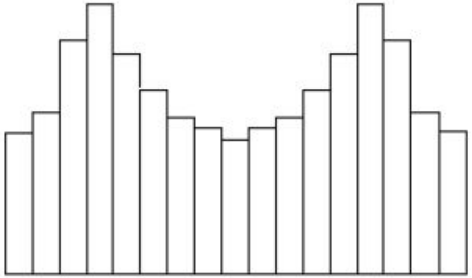
Distribuição com assimetria negativa

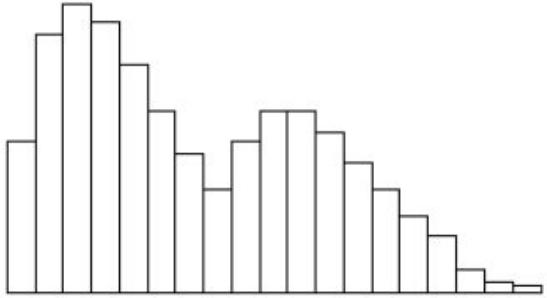




Ele parece um barranco, como se as frequências despencassem.

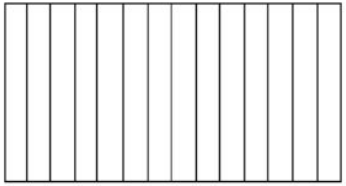
Neste caso, os valores mais altos concentram-se à direita ou à esquerda, com seu valor médio fora do centro.





Este tipo de gráfico chamado de bimodal é característico por dois picos mais altos em regiões diferentes.

Isso ocorre sempre quando há uma mistura de dados. Quando temos uma variável coletada para dois grupos que se comportam completamente diferentes.



Uniform (no mode)

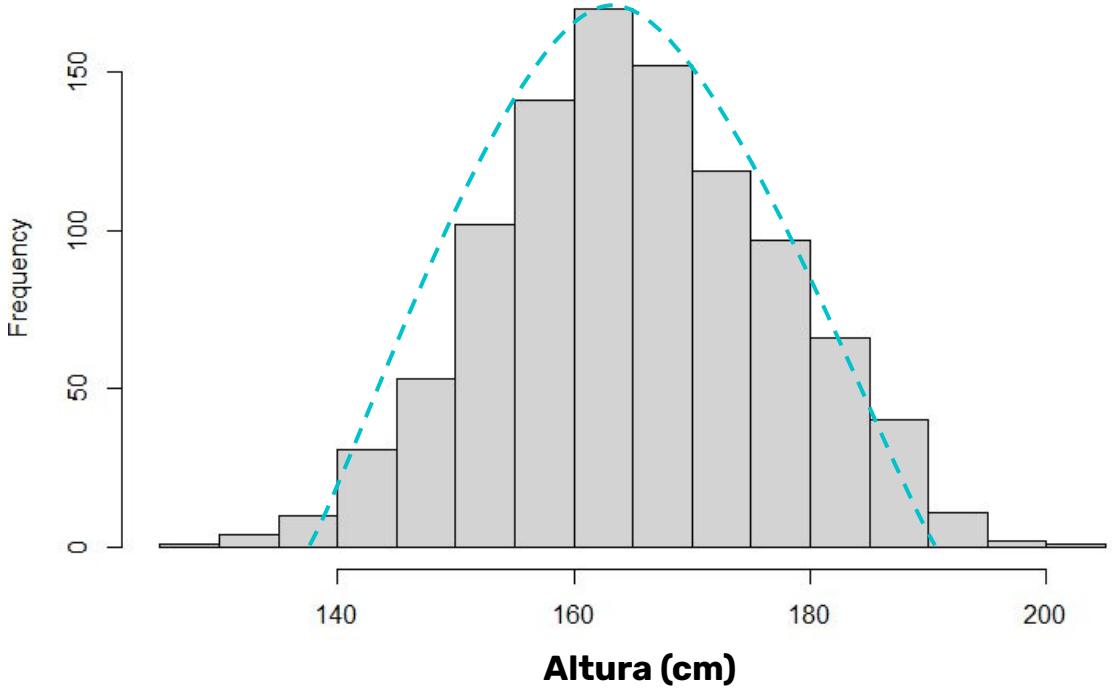
Este histograma é característico por apresentar frequências de nível equivalente bem próximas uma das outras.

ALGUMA DÚVIDA ATÉ AQUI?

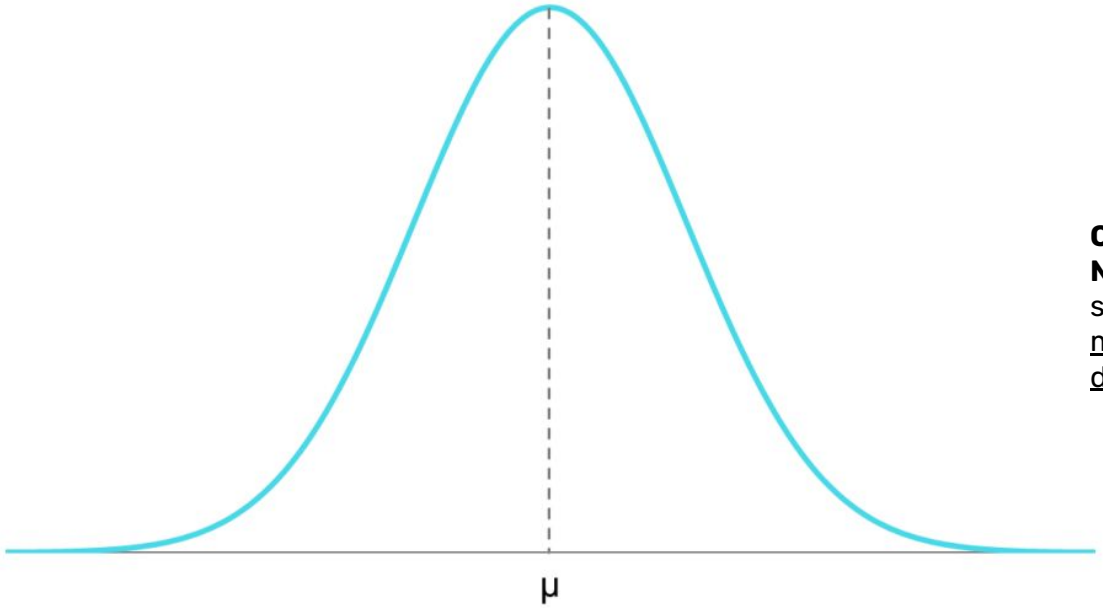
Distribuições

Distribuição normal (Gaussiana)

A **distribuição Normal** é a mais familiar das distribuições de probabilidade e também uma das mais importantes em estatística. É considerada uma distribuição normal, se a v**ariável aleatória contínua** tem uma distribuição com um gráfico simétrico em forma de sino:



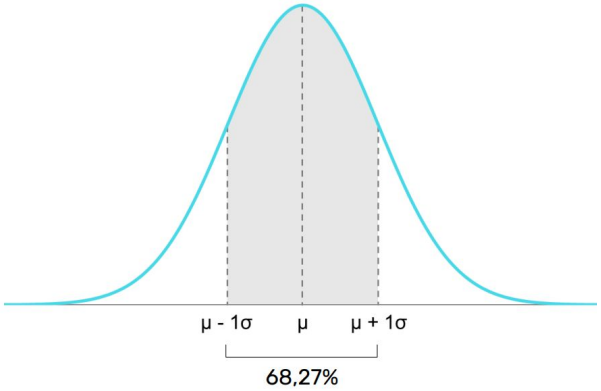
Distribuição normal



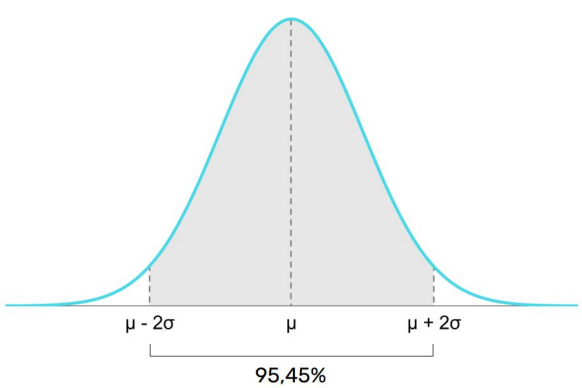
$\mu =$ média populacional

Os dados que possuem distribuição Normal (simetricamente distribuídos) seguem um padrão no qual o valor da maioria dos pontos dos dados fica dentro de três desvios padrão da média.

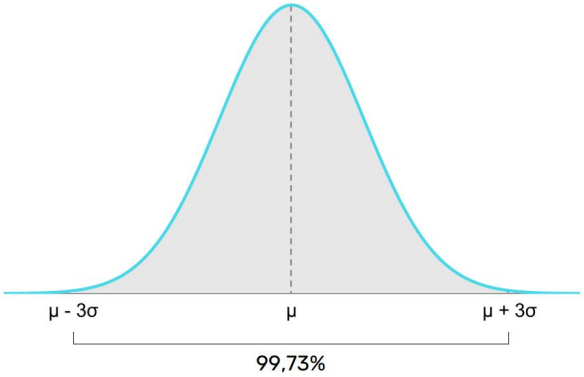
Distribuição normal



Cerca de 68,27% de todos os valores ficam a um desvio padrão da média.



Cerca de 95,45% de todos os valores ficam a dois desvios padrão da média.

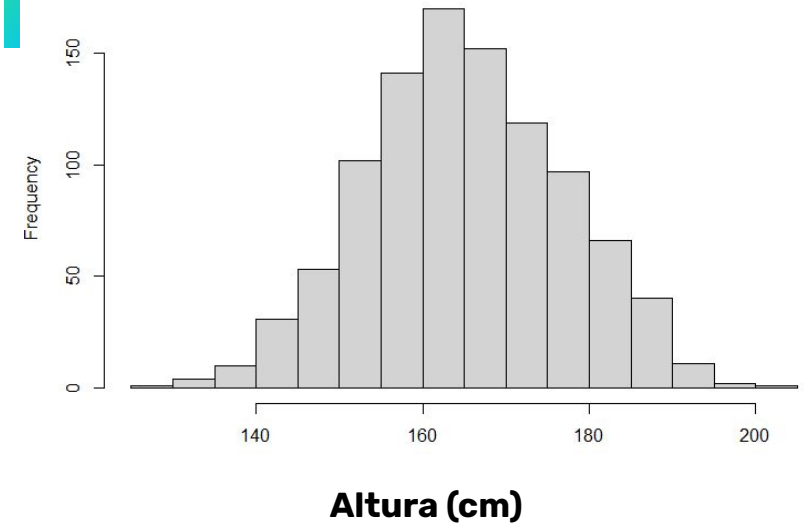


Cerca de 99,73% de todos os valores ficam a dois desvios padrão da média.

$$\mu = \bar{x} \qquad \sigma = s$$

Exemplo

Média: 165 cm
Desvio-padrão: 12 cm



1 desvio da média:

- 165 - (1x12) = 153 cm
- 165 + (1x12) = 177 cm

68,27% das pessoas têm alturas entre 153 e 177 cm.

2 desvios da média:

- 165 - (2x12) = 141 cm
- 165 + (2x12) = 189 cm

95,94% das pessoas têm alturas entre 141 e 189 cm.

3 desvios da média:

- 165 - (3x12) = 129 cm
- 165 + (3x12) = 201 cm

99,7% das pessoas têm alturas entre 129 e 201 cm.

Mão na Massa

Ao fazer um teste de QI com uma amostra, você identificou que o escore de QI segue uma distribuição Normal com média de 100 e o desvio padrão de 15.

O que podemos afirmar sobre 95% das pessoas?

Mão na Massa (Gabarito)

Ao fazer um teste de QI com uma amostra, você identificou que o escore de QI segue uma distribuição Normal com média de 100 e o desvio padrão de 15.

O que podemos afirmar sobre 95% das pessoas?

Média: 100
Desvio-padrão: 15

1 desvio da média:

- 100 - (1x15) = 85
- 100 + (1x15) = 115

68,27% das pessoas têm o QI entre 85 e 115.

2 desvios da média:

- 100 - (2x15) = 70
- 100 + (2x15) = 130

95,94% das pessoas têm o QI entre 70 e 130.

3 desvios da média:

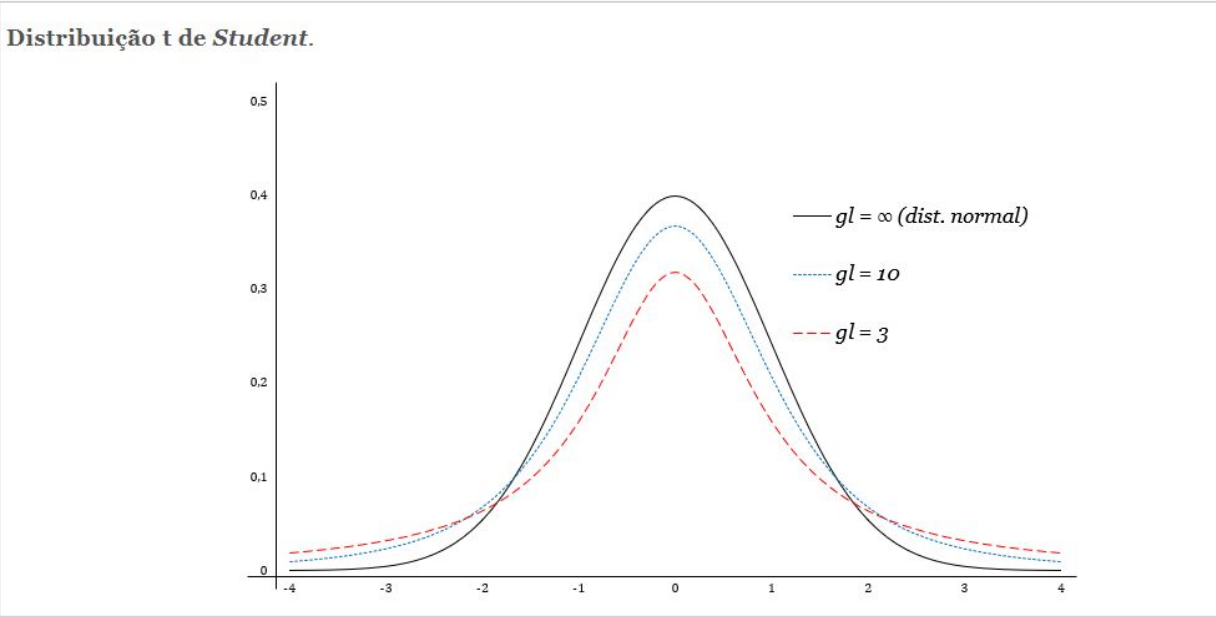
- 100 - (3x15) = 55
- 100 + (3x15) = 145

99,7% das pessoas têm o QI entre 55 e 145.

Distribuição de t-Student

É muito semelhante à distribuição normal também em forma de sino e simétrica em relação a média.
Utilizamos essa distribuição quando temos amostras pequenas e a variância da população é desconhecida.

Como consequência, temos curvas com caudas mais pesadas, ou seja, os extremos têm probabilidades maiores.



Distribuição de Binomial

A **distribuição binomial** nos permite lidar com circunstâncias nas quais os resultados pertencem a duas categorias relevantes, tais como **aceitável/defeituoso** ou **sobreviveu/morreu**.

Uma distribuição de probabilidade binomial resulta de um experimento que satisfaz os seguintes requisitos:

- O experimento tem um número fixo de tentativas.
- As tentativas tem que ser independentes. (O resultado de qualquer tentativa não afeta as probabilidades nas outras tentativas).
- Cada tentativa deve ter todos os resultados classificados em duas categorias (em geral, chamadas de **sucesso e fracasso**).
- A probabilidade de um sucesso permanece constante em todas as tentativas.

Exemplos :

- O número de **caras** em n **tentativas** independentes.
- O número de **itens defeituosos** em n **itens produzidos** independentes.

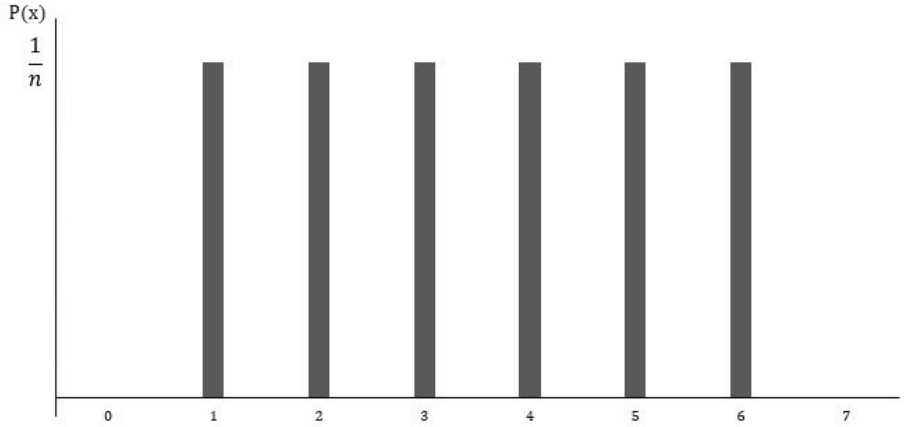
Nota: A palavra Sucesso é arbitrária e não representa necessariamente algo bom.
Qualquer uma das suas duas categorias pode ser chamada de sucesso (S), desde que sua probabilidade seja identificada como p

Distribuição Uniforme (discreta)

Uma variável possui distribuição uniforme quando os valores em um intervalo, possuem a mesma chance de ocorrer, ou seja, entre *n* valores possíveis, cada um recebe a probabilidade 1/*n* de ocorrer.

Exemplo:

Um exemplo é um lançamento de um dado: os valores possíveis são {1,2,3,4,5,6} e cada valor possui probabilidade de $\frac{1}{6}$.



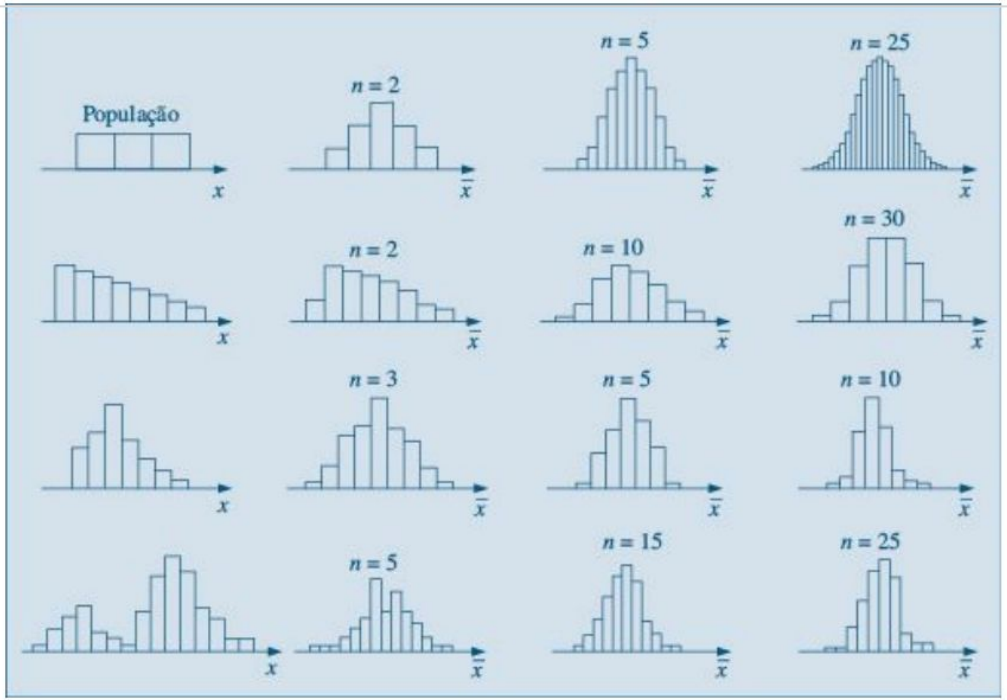
Distribuição qui-quadrado

Derivação da Gama, que é derivação da distribuição exponencial.

Teorema do limite central

Quando o tamanho da amostra aumenta, independentemente da forma da distribuição da população, a distribuição amostral de $\bar{X}$ aproxima-se cada vez mais de uma distribuição normal.

Esse resultado, fundamental na teoria da Inferência Estatística, é conhecido como Teorema Limite Central (TLC).



DÚVIDAS FINAIS

COMO FOI?


30° 45° 60°
sin 1/2 √2/2 √3/2
cos √3/2 √2/2 1/2
tan √3/3 1 √3
2x
60°
x
30°
x√3
45°
s√2
s
45°


AH! AGORA EU ENTENDI

Boa noite
E ATÉ A PRÓXIMA!